Digitization of Konkani Texts, and their Transliteration: An Initiative towards Preservation of a Language Culture

Ms. Palia Tukaram Gaonkar[1] and Dr. Andre Rafael Fernandes[2]

[1] Doctoral Research Scholar, Department of English, Goa University, Goa, India

eng.palia@unigoa.ac.in

[2] Associate Professor, Department of English, Goa University, Goa, India

rafael@unigoa.ac.in

www.unigoa.ac.in

Abstract. Konkani is the official language of the state of Goa, located on the western coast of India. This language has faced many political threats such as four hundred and fifty years of Portuguese colonization and contention with Marathi in order to be recognised as the official language, post-Liberation in 1961. It finally entered the Eighth Schedule of the Constitution of India in 1992. These hardships have diversified the nature of Konkani. It is spoken in several dialects in the state of Goa and elsewhere, and is written in five different scripts, owing to the migration of Konkani people from Goa over the centuries. There are Konkani communities in the neighbouring states of Karnataka, Kerala and Maharashtra, which are heavily influenced by the dominant local culture. Hence, Konkani is written in Devanagari, Roman, Kannada, Malayalam and Perso-Arabic scripts.

This phenomenon creates a linguistic and literary gap in the community of Konkani speakers. Persistent efforts to bridge the gap have been made, and one of them is by taking the assistance of technology. The World Konkani Centre situated in Mangalore, Karnataka, has developed a transliteration tool, *Konkanverter*, which transliterates Unicode text between four of the five writing systems of Konkani, namely Devanagari, Roman, Kannada and Malayalam. This paper reports the attempts to digitize, and transliterate an available performance play text in Konkani, from one script to another. It also explores digitization as a way of preserving Konkani texts in multi-script formats.

Keywords: Konkani, Digitization, Transliteration

1 Introduction

The official language of Goa, Konkani, also spoken in pockets of Karnataka, Kerala and Maharashtra, showed a negative growth rate according to the Census of 2011[1]. Of the total population of India, 0.19% of the population claimed it to be their mother tongue; but what is more alarming to notice is that the number of native speakers decreased from 24,89,015 to 22,56,502, with a growth rate of -9.34% [1].

In addition to this apprehension, The Telegraph, a noteworthy Indian national daily published the interview of language expert Professor G N Devy in August 2017 after he had published several volumes of the People’s Linguistic Survey of India, wherein he observed that, “Nearly 400 of India's 850-odd languages face the threat of extinction because of an erosion of traditional jobs that is fuelling migration to cities,” and that “the languages spoken in the coastal areas will be the worst-affected” [2].

Goa has been through the perils of colonization, mechanization, migration and so forth. In addition, the Konkani language carries a linguistic peculiarity of being written in five different scripts, with two or more scripts being used in the same geographical region. The literature created in these five scripts remains restricted to readers who are familiar with the scripts. Their geographical location decides their readability of the scripts. Therefore, there will be very few who can read in three scripts, let alone four or five.

However, it is remarkable that its cultural and linguistic identity has been preserved by its people quite fervently, despite the grave threats that appeared time and again to erase it. The Konkani people carried their language, their deities and their culture to the places they migrated to, and planted their cultural heritage in alien soil. However, in a scenario where technological tools greet the end-users in global languages, a minority language can easily be bypassed. Although close to nine of the twenty-two scheduled languages recognized by the Constitution of India are featured by Google, so far, Konkani has not been one of them. However, under Google Indic Keyboard, Konkani-English interface is provided, with a limited corpus, in Android smart-phone keyboards.

If a small language like Konkani is to survive the invasion of technology and its mission to standardize communication, it needs to adopt technology to preserve its diversity, and subsequently its culture and heritage. This study aims to explore the possibility of having a cross-orthographical readership of Konkani using the transliteration tool known as *Konkanverter*, which will not only bridge the orthographic disparity and increase readership and production of literature, but also contribute to the creation of Konkani digital archive.

2 Literature Review

Due to the migration of people from Goa over the centuries, Konkani diasporic communities exist in Karnataka, Kerala, and Maharashtra. It is not surprising that their spoken dialect of Konkani carries the flavour of the regional languages in the respective states, which are Kannada, Malayalam and Marathi respectively. Hence the local scripts have been adopted by these Konkani communities in the respective regions.

The two dominant scripts of Konkani in Goa are Devanagari and Roman or Romi as it is colloquially known. Devanagari orthography is considered the official one; Romi is used by select weeklies/ magazines and *Tiatr* (drama) scripts.

Rajan terms this phenomenon of multiple scripts as “synchronic trigraphia” and goes on to elucidate it as “a major issue of political contention inside the community, each group favouring the usage of a particular script as the official orthography. Different orthographic communities exist in isolation with minimal interaction and with its [sic] own literary tradition, as very few people are fluent in multiple orthographies” [3].

Although, these communities may have elementary proficiency in reading in the other writing system, this multiplicity of scripts creates an obstacle for wider reading. Hence, literature and language both become less accessible to the native speakers of a single region: “This disparity in scripts creates intra-lingual barriers which make writings in Konkani inaccessible even to a native Konkani speaker who has limited or no knowledge of all these five scripts” [4].

Rajan suggests that a “statistical machine transliteration engine with reasonable accuracy would greatly enable cohesion and interaction among the greater linguistic community. Facilitating the usage of multiple scripts would also encourage more linguistic diversity among the community” [3]. Rajan further goes on to present the development of such a tool, Konkanverter (http://konkanverter.com/) [5], which has been extensively used for this study.

3 Methodology

The methodology used in this study is exploratory in nature. As Dudovskiy puts it, “Exploratory research, as the name implies, intends merely to explore the research questions and does not intend to offer final and conclusive solutions to existing problems” [6]. Exploratory research provides insights into a given situation, and produces qualitative research that becomes the basis of further research in the specified study area. Exploration involves planning (deciding upon the specific research questions), exploration (data collection), reflection and analysis (data interpretations and observations) [7].

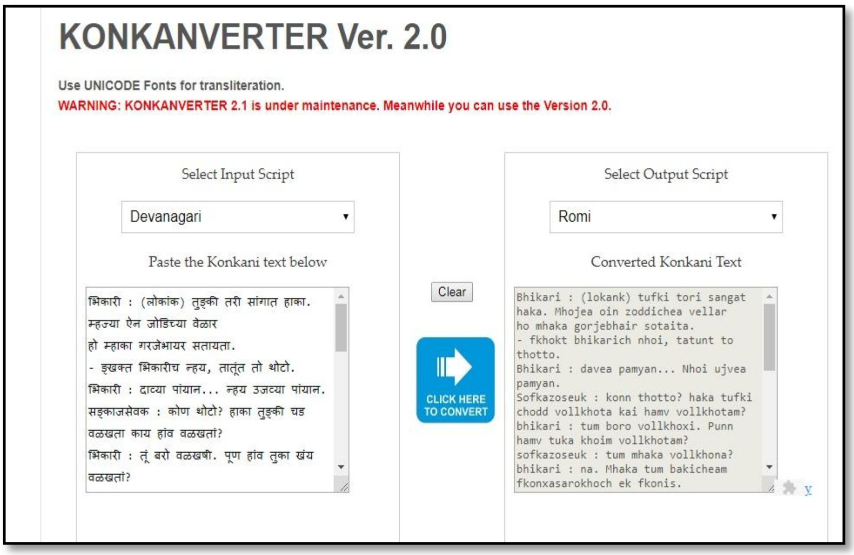
In the context of this study, exploratory research helps discuss benefits, and challenges that come in the way, of solving the major research question, i.e. whether digitization can pave a way to language preservation. The present study adheres to this methodology in a way that is applicable to the discipline. The study serves to explore the possibilities of using technology to safeguard the linguistic-cultural heritage of a certain language community, and paves way for further research in this area. The data and analysis presented in this study are first-hand; revision can be taken up as part of the next level of research in the same area.

4 Transliteration of Konkani Play “Shree Vichitrachi Jatra”

World Konkani Centre [8], Mangalore, along with computer scientist Vinodh Rajan, has introduced a “finite state transducer based transliteration engine” [3] known as Konkanverter. This tool provides transliteration in four popular orthographies of Konkani, namely Devanagari, Romi, Kannada and Malayalam [5].

The present research involved transliteration of a well-known Konkani play “Shree Vichitrachi Jatra”, written by Pundalik Naik. This play belongs to national award winning drama collection, *Chourang* (1982) **[**9**]**. The selection of the text was based on its popularity and critical acclaim; its availability in the digital format also facilitated its conversion.

The digital copy of the play was obtained in Devanagari. This text had lexical errors in it, although not largely phonetic ones. Furthermore, it was found that the font used was not in Unicode, and since Konkanverter only accepted fonts in Unicode, an intermediate software tool known as Baraha FontConvert [10] was used. Since FontConvert has a limit on the amount of input text, the process had to be undertaken through several sessions. After converting the font from “Shree-Dev-0709<==>BRH Devanagari”, the text output in Unicode was then pasted in the left-hand side text input box in Konkanverter. After clicking on the conversion tab, the transliterated text was obtained in the right-hand side text output box (See Figure 1). This text was then pasted into Notepad.



KONKANVERTER Ver. 2.0

Use UNICODE Fonts for transliteration.

WARNING: KONKANVERTER 2.1 is under maintenance. Meanwhile you can use the Version 2.0.

Select Input Script: Devanagari

Paste the Konkani text below

भिकारी : (लोकांक) तुड्की तरी सांगात हाका.
म्हज्या ऐन जोडिच्या वेळार
हो म्हाका गरजेभायर सतायता.
- इ्खक्त भिकारीच न्हय, तातूंत तो थोटो.
भिकारी : दाव्या पांयान... न्हय उजव्या पांयान.
सड्काजसेवक : कोण थोटो? हाका तुड्की चड
वळखता काय हांव वळखतां?
भिकारी : तूं बरो वळखषी. पूण हांव तुका खंय
वळखतां?

Select Output Script: Romi

Converted Konkani Text

```
Bhikari : (lokank) tufki tori sangat
haka. Mhojea oin zoddichea vellar
ho mhaka gorjebhair sotaita.
- fkhokt bhikarich nhoi, tatunt to
thotto.
Bhikari : davea pamyan... Nhoi ujvea
pamyan.
Sofkazoseuk : konn thotto? haka tufki
chodd vollkhota kai hamv vollkhotam?
bhikari : tum boro vollkhoxi. Punn
hamv tuka khoim vollkhotam?
sofkazoseuk : tum mhaka vollkhona?
bhikari : na. Mhaka tum bakicheam
fkonxasarokhoch ek fkonis.
```

Fig.1. Screenshot of Konkanverter converting text of “Shree Vichitrachi Jatra”.

It was observed that the source text font (Shree-Dev) could not be accurately converted to the Unicode format, and hence some of the letters were wrongly converted. These errors were carried forward in the transliteration tool as well. Table 1 below indicates the source text (ST), followed by the text converted into Unicode (UT), which is further followed by the transliterated text (TT). The underlined error in the UT is phonetically unreadable (इ्ख), which is transferred to the TT. The expected output is also displayed in the fourth row of the table (see Table 1).

Table 1. Table indicating transfer of errors in text conversion.

Text Type	Text
ST	कट्टी, झोळी, पिषवी सगळें एकूच तर तूं ताचे पिषवेंत एक फुटकी कवडी वडयना आनी वयल्यान ताचीं फकाणां करता?
UT	कट्टी, झोळी, पिषवी सगळें एकूच तर तूं ताचे पिषवेंत एक ङ्खुटकी कवडी वडयना आनी वयल्यान ताचीं ङ्खकाणां करता?
TT	kott'tti, zholli, pixvi sogllem ekuch tor tum tache pixvent ek fkhuttki kovddi voddoina ani voilean tachim fkhokannam korta?
Expected Output	Kott'tti, zholli, pixvi sogllem ekuch tor tum tache pixvent ek futtki kouddi voddoina ani voilean tachim fokannam korta?

As given in Table 1, one can clearly notice that the addition of 'kh' syllable happens at the first stage of conversion, i.e. when the digital source text is converted into Unicode.

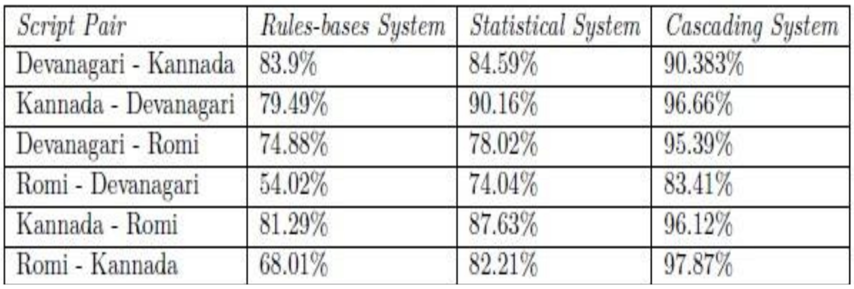
Table 2 explores ten random errors found in the transliterated text, and discusses the consistency of errors in the output. It shows column-wise progression, from source text, copied and pasted into font conversion tool (see Table 2 column 'Source Font') to receive the Unicode output (see Table 2 column 'Unicode Output Text'). This Unicode output text was copied and pasted into Konkanverter to give its transliteration (see Table 2 column 'Transliterated Output Text'). The content of the last column of Table 2, i.e. 'Expected Transliteration', is a result of directly typing the corresponding words of the column 'Source Text' of Table 2 in Unicode into Konkanverter input text box, and the output received is not influenced by font conversion and hence is the expected machine transliteration.

Table 2. Table showing examples of the subsequent text error-transfer in font conversion and machine transliteration.

Sr, No	Source Text	Source Font	Unicode Output Text	Transliterated Output Text	Expected Transliteration
1	फक्त	’$ŠV	ङ्खक्त	fkhokt	fokt
2	समाजसेवक	g‘mOgodH$	सङ्काजसेवक	sofkazsevok	somazoseuk
3	मनीस	‘Zrg	ङ्कनीस	fkonis	monis
4	आसलो	Amgcm	आसला	asla	aslo
5	म्हळ्यार	åhiçma	म्हळार	mhollar	mhollear
6	कोणतरी	H$mUo Var	काणे तरी	konne tori	konntori
7	बोमाड्यान	~mo‘mS>çmZ	बोङ्काडान	bofkaddan	bomaddean
8	गाड्याक	JmS>çmH$	गाडाक	gaddak	gaddeak
9	वाट्याचें	dmQ>çmM|	वाटाचें	vattachem	vatteachem
10	तुळा	Viw m	तळु	tollu	tulla

As given in Table 2, one can observe that the letters फ (fa), म (ma) seem to have been converted erroneously throughout the text; other major erroneous conversions are the joint half letters such as म्हळ्यार- म्हळार; गाड्याक- गाडाक; वाट्याचें- वाटाचें. Some seem to be random errors (UT: आसला, तळु, काणेतरी) which were not repetitive.

The entire play was thus transliterated into Romi, and errors like the above (see Table 2) had to be dealt with manually. At the time of writing this paper, it is unknown whether these discrepancies would get transferred into Kannada or Malayalam, as the present researchers are not familiar with the spoken dialect or the scripts of these regions. Rajan [3] gives the table for accuracy of the transliteration as follows:



Script Pair	*Rules-bases System*	*Statistical System*	*Cascading System*
Devanagari - Kannada	83.9%	84.59%	90.383%
Kannada - Devanagari	79.49%	90.16%	96.66%
Devanagari - Romi	74.88%	78.02%	95.39%
Romi - Devanagari	54.02%	74.04%	83.41%
Kannada - Romi	81.29%	87.63%	96.12%
Romi - Kannada	68.01%	82.21%	97.87%

Fig.2. Table indicating accuracy of three different systems used by Konkanverter transliteration tool (Source**:** Rajan, V. [3] p. 18).

The above transliteration accuracy rates are to be considered in the case of a text which is lexically accurate. A limitation of this study was that the source text was not proofread for spelling errors, which led to rise in the inaccuracy of transliteration. Therefore, care has to be taken by Konkani language experts to have it perfectly proofread after digitization, and also have the font converted to Unicode, which will ensure faithful transliteration.

For more accuracy to be obtained in translation, ST should be proofread and converted into Unicode, which will render the intermediate font conversion redundant. With only one conversion engine between the input and the output text, greater accuracy will easily be achieved.

Such an analysis of Konkanverter is incomplete without references to Google's contribution to the field of transliteration. Transliteration is featured as one of *Google Input Tools* [11], which provides on-the-fly options to Roman rendition of a Devanagari word. However, a Devanagari rendition in Unicode does not get converted to its Roman equivalent. Hence, transliteration takes place only in one way. Moreover, the Devanagari script provision is only for Hindi and Marathi languages, Konkani is not featured in the options. As far as Google Indic Keyboard on Android smart-phones is concerned, Konkani type font is available in Roman and Devanagari, although transliteration takes place only from Roman to Devanagari Konkani.

5 Conclusion

The acceptance of only the Unicode text limits the usage of Konkanverter, as it discourages input of any other kind, such as optically recognized characters in general and voice input. Voice input is further deterred by not having the text on-the-fly.

Nevertheless, Konkanverter pushes the boundaries of learning of different scripts of the same language, and makes multi-script archiving possible; it not only bridges the gap between the two dominant orthographic Konkani communities, but establishes a kind of digital footprint for a language which could face the risk of becoming endangered. Such a tool can become an inspiration for minority languages across the globe to enhance their digital presence, and hence safeguard their identity in the digital revolution.

References

1. Ministry of Home Affairs, Government of India. 2013. Statement - 7 Growth Of Scheduled Languages - 1971, 1981, 1991, 2001 and 2011.Office of the Registrar General & Census Commissioner, India. http://www.censusindia.gov.in/2011Census/Language_MTs.html, last accessed 2018/10/10.
2. Mohanty, B. K.: Extinction alert on 400 languages. The Telegraph: Online edition. https://www.telegraphindia.com/india/extinction-alert-on-400 languages/cid/1521283, last accessed 2018/10/05.
3. Rajan, V.: Konkanverter - A Finite State Transducer based Statistical Machine Transliteration Engine Konkani Language. Proceedings of the 5th Workshop on South and Southeast Asian NLP, 25th International Conference on Computational Linguistics, pp. 11–19. Ireland (2014).
4. Gaonkar, P., Fernandes A. R.: ICT in Language Teaching Analysis of Select Software. Proceedings of the International Conference on Trends and Innovations in Language Teaching, pp .221. Sathyabama University, Chennai (2014).
5. Konkanverter. World Konkani Centre, Version 2.0. World Konkani Centre Mangalore. http://konkanverter.com/, last accessed 2014/12/18.
6. Dudovskiy, J.: Exploratory Research. Research Methodology. https://research-methodology.net/research methodology/research-design/exploratory-research/#_ftn2, last accessed 2018/10/02
7. Smith, R., Rebolledo, P.: A Handbook for Exploratory Action Research. British Council (2018).
8. World Konkani Centre. World Konkani Centre Mangalore. http://vishwakonkani.org/, last accessed 2018/10/19
9. Naik, P. Chourang. Apurbai Prakashan, Goa (1982).
10. FontConvert. Baraha Indian Language Software. Baraha Software (2014).
11. Google Input Tools. Google. https://www.google.com/inputtools/try/, last accessed 2019/02/01.